Journal of Culture, Society and Development- An Open Access International Journal
Vol.2 2013

An Alternative Approach to Estimation of Population Mean in Two-Stage Sampling

Ravendra Singh[1]*, P.C. Gupta[1], Sarla Pareek[1], Gajendra K. Vishwakarma[2]
1.Department of Mathematics & Statistics,Banasthali University, Jaipur- 304022, Rajasthan, India
ravendrasingh84@gmail.com
2.Department of Applied Mathematics, Indian School of Mines, Dhanbad - 826004, Jharkhand, India
vishwagk@rediffmail.com

Abstract
This paper considers the problem of estimating the population mean in two-stage sampling with unequal first stage unit (fsu) for ratio, product and regression estimators when the population mean of the auxiliary variable is not known in advance. The ratio, product and regression estimators are suggested and studied their properties. It is shown that under certain conditions the suggested estimators are more efficient than than Sukhatme et al (1984) estimators. Numerical illustration is given in support of present study.
Keywords: Auxiliary variate, study variate, mean squared error, two-stage sampling, ratio estimator, product estimator, regression estimator.
Mathematics Subject Classification: 62D05.

1. Introduction

In survey sampling, the use of auxiliary information can increase the precision of an estimator when study variable y is highly correlated with the auxiliary variable x. but in several practical situations, instead of existence of auxiliary variables there exists some auxiliary attributes, which are highly correlated with study variable y, such as (i) Amount of milk produced and a particular breed of cow. (ii) Yield of wheat crop and a particular variety of wheat etc. In such situations, taking the advantage of point bi-serial correlation between the study variable and the auxiliary attribute, the estimators of parameters of interest can be constructed by using prior knowledge of the parameters of auxiliary attribute.

The problem of the estimation of population parameters like mean, variance, and ratio of two population means are common in agriculture, economics, medicine, and population studies. The use of auxiliary information has been applied for improving the efficiencies of the estimators of population parameter(*s*) irrespective of sampling design. Ratio, product and regression methods of estimation are good examples in this context. Cochran (1940) used auxiliary information at the estimation stage and proposed a ratio estimator for the population mean. A ratio estimator is preferred when the correlation coefficient between the study variate and the auxiliary variate is positive. Robson (1957) defined a product estimator that was revisited by Murthy (1964). The product estimator is used when the correlation between the study variate and the auxiliary is negative.

The use of auxiliary information has to a fairly large extent proved to be effective approach in increasing the precision of estimates in survey sampling. Many estimators of the finite population parameter have been constructed using auxiliary information. Information on variables correlated with the main variable under study is popularly known as auxiliary information which may be fruitfully utilized either at planning stage or design stage or at the estimation stage to arrive at an improved estimate compared to those not utilizing it. The use of auxiliary information dates back to 1820 with a comprehensive theory provided by Cochran (1977). Recently, several authors like Samiuddin and Hanif (2007), Kadilar and Cingi (2004, 2005), Pradhan (2005), Sahoo and Panda (1997, 1999) and Upadhyaya and Singh (1999) have worked on the use off auxiliary information in survey sampling.

In some cases information on auxiliary variables may be readily available on all units in the population; however, this is not always the case in certain practical situations. Hence, we rely on taking a fairly large sample from the N unit in the population in order to estimate the population mean of the auxiliary variable(s).

Since it is assumed that the population mean $\overline{X}$, of the auxiliary variable, is unknown, we first select a preliminary large sample of size n' from N units in the population by simple random sampling without replacement in order to provide an estimate of $\overline{X}$. Let U be a finite population partitioned into n' First Stage Units (FSU) denoted by $U_1, U_2, ..., U_I, ...U_{n'}$ such that the number of Second Stage Units (SSU) in U_i is M_i. Let y and x be the variable under study and auxiliary variable taking the values y_{ij} and x_{ij} respectively, for the j^{th} SSU on the unit U_i $\left(i = 1,2,...,n'; j = 1,2,...,M_i\right)$.

$$\overline{P}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} P_{ij} \; ; \; \overline{P} = \frac{1}{n'}\sum_{i=1}^{n'} \overline{P}_i \; ; \; P = X \; , \; Y \; .$$

Assume that a sample s of n FSU's is drawn from U and then a sample s_i of m_i SSU's from the i^{th} selected FSU from U_i using simple random sampling without replacement.

$$\bar{p}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} p_{ij}\ ;\ \bar{p} = \frac{1}{n}\sum_{i=1}^{n}\bar{p}_i\ ;\ p = x,\ y\ .$$

When $\bar{X}$ is known, the classical two-stage ratio, product and regression estimators of $\bar{Y}$ with unequal FSU and their corresponding approximate Mean Square Error (MSE) has been given by Sukhatme *et al* (1984) and are as presented below:

Ratio estimator:

$$\bar{y}_1 = \bar{y}\frac{\bar{X}}{\bar{x}} \tag{1}$$

$$MSE(\bar{y}_1) = \lambda S_{1R}^2 + \sum_{i=1}^{N}\lambda_{2i}S_{2Ri}^2$$

Product estimator

$$\bar{y}_2 = \bar{y}\frac{\bar{x}}{\bar{X}} \tag{2}$$

$$MSE(\bar{y}_2) = \lambda S_{1P}^2 + \sum_{i=1}^{N}\lambda_{2i}S_{2Pi}^2$$

Regression estimator

$$\bar{y}_3 = \bar{y} + \beta(\bar{X} - \bar{x}) \tag{3}$$

$$MSE(\bar{y}_2) = \lambda S_{1B}^2 + \sum_{i=1}^{N}\lambda_{2i}S_{2Bi}^2$$

where $\lambda = \frac{1-f}{n}; f = \frac{n}{N};\ \lambda_{2i} = \frac{M_i^2(1-f_{2i})}{nNm_i};\ f_{2i} = \frac{m_i}{M_i};$

$S_{1R}^2 = S_{1y}^2 + R^2S_{1x}^2 - 2RS_{1xy};\ S_{1P}^2 = S_{1y}^2 + R^2S_{1x}^2 + 2RS_{1xy};\ S_{1B}^2 = S_{1y}^2 + B^2S_{1x}^2 - 2BS_{1xy};$

$S_{2Ri}^2 = S_{2yi}^2 + R^2S_{2xi}^2 - 2RS_{2xyi};\ S_{2Pi}^2 = S_{2yi}^2 + R^2S_{2xi}^2 + 2RS_{2xyi};$

$S_{2Bi}^2 = S_{2yi}^2 + B^2S_{2xi}^2 - 2BS_{2xyi};\ S_{1p}^2 = \frac{1}{n'-1}\sum_{i=1}^{n'}(\bar{P}_i - \bar{P})^2;\ S_{2Pi}^2 = \frac{1}{n'-1}\sum_{i=1}^{n'}(P_{ij} - \bar{P}_i)^2;$

$$S_{1PQ} = \frac{1}{n'-1}\sum_{i=1}^{n'}(\bar{P}_i - \bar{P})(\bar{Q}_i - \bar{Q});$$

$$S_{2PQi} = \frac{1}{n'-1}\sum_{i=1}^{n'}(P_{ij} - \bar{P}_i)(Q_{ij} - \bar{Q}_i);\ P = x,\ y;\ Q = x,\ y \text{ and } P \neq Q.$$

S_{1P}^2 and S_{2Pi}^2 are the variance among FSU means and variance among subunits for the i^{th} FSU while S_{1PQ} and S_{2PQi} are their corresponding co-variances.

Sahoo and Panda (1997, 1999) described class of estimators in two stage sampling with varying probabilities and later examined family of estimators using auxiliary information in two-stage sampling. Scott and Smith (1967) and Chaudhuri and Stenger (1992) described the prediction criterion for two-stage sampling while Hossain and Ahmed (2001) examined the class of predictive estimators in two-stage sampling using auxiliary information.

2. Suggested Estimators

RATIO ESTIMATOR

Under the assumptions above, the estimator $\bar{y}_1$ will take the form

$$y_1^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' \tag{4}$$

If we express (4) in terms of d's

$$\bar{y}_1^* = Y(1+d_o)(1+d_2)(1+d_1)^{-1} \quad (5)$$

Expanding the right hand side of (5) neglecting terms involving power of this greater than two, we have

$$\bar{y}_1^* = \bar{Y}\left(1+d_0-d_1+d_2+d_0d_2-d_0d_1-d_1d_2+d_1^2\right)$$

$$\bar{y}_1^* - \bar{Y} = \bar{Y}(d_0-d_1+d_2+d_0d_2-d_0d_1-d_1d_2+d_1^2) \quad (6)$$

Squaring both sides of (6) and neglecting terms involving powers greater than two, we have

$$\left(\bar{y}_1^* - \bar{Y}\right)^2 = \bar{Y}^2\left(d_0^2+d_1^2+d_2^2+2d_0d_2-2d_0d_1-2d_1d_2\right) \quad (7)$$

By taking expectations of both sides by (7) and simplifying we get the MSE of as

$$MSE\left(\bar{y}_1^*\right) = \lambda' S_{1R}^2 + \sum_{i=1}^{n'} \lambda_{2i} S_{2Ri}^2 - \lambda_3 S_{1R'}^2 - \sum_{i=1}^{n'} \lambda_{3i} S_{2Ri}^2 \quad (8)$$

where $S_{1R'}^2 = R^2 S_{1x}^2 - 2RS_{1xy}$; $S_{2Ri'}^2 = R^2 S_{2xi}^2 - 2RS_{2xyi}$

PRODUCT ESTIMATOR is defined as

$$\bar{y}_2^* = \bar{y}\frac{\bar{x}^*}{\bar{x}'}$$

Following the procedure of section 2 it is easy to verify that the MSE of $\bar{y}_2^*$ is

$$MSE\left(\bar{y}_2^*\right) = \lambda' S_{1P}^2 + \sum_{i=1}^{n'} \lambda_{2i} S_{2Pi}^2 - \lambda_3 S_{1P'}^2 - \sum_{i=1}^{n'} \lambda_{3i} S_{2Pi}^2 \quad (9)$$

where, $S_{1P'}^2 = R^2 S_{1x}^2 + 2RS_{1xy}$; $S_{2Pi'}^2 = R^2 S_{2xi}^2 + 2RS_{2xyi}$

REGRESSION ESTIMATOR is defined by

$$\bar{y}_3^* = \bar{y}^* + \hat{\beta}^*(\bar{x}' - \bar{x}^*)$$

where $\hat{\beta}^* = \frac{s_{xy}^*}{s_x^{*2}}$ is an estimator of $\beta = \frac{S_{xy}}{S_x^2}$. It is easy to verify that s_{xy}^* and s_x^{*2} are unbiased for S_{xy} and S_x^2 respectively.

Let $\bar{y}^* = \bar{Y}\left(1+d_0\right)$, $\bar{x}^* = \bar{X}\left(1+d_1\right)$, $\bar{x}' = \bar{X}\left(1+d_2\right)$

such that $E\left(d_{P^*}\right) = 0$; $E\left(d_{P^*}^2\right) = \frac{V\left(\bar{P}^*\right)}{\bar{P}^{*2}}$; $E\left(d_{P^*}, d_{Q^*}\right) = \frac{Cov\left(\bar{P}^*, \bar{Q}^*\right)}{\bar{P}^*\bar{Q}^*}$; $E\left(d_2^2\right) = \frac{V\left(\bar{x}'\right)}{\bar{X}^2}$;

$$E\left(d_{P^*}d_2\right) = \frac{Cov\left(\bar{P}^*, \bar{x}'\right)}{\bar{P}^*\bar{X}}$$

Under TSS with unequal FSU

$$V\left(\bar{P}^*\right) = \lambda' S_{1P^*}^2 + \sum_{i=1}^{n'} \lambda_{2i} S_{2P^*i}^2 \text{ ; } Cov\left(\bar{P}^*, \bar{Q}^*\right) = \lambda' S_{1P^*Q^*} + \sum_{i=1}^{n'} \lambda_{2i} S_{2P^*Q^*i} \text{ ; } \lambda' = \frac{1-f_1}{n} \text{ ;}$$

$$f_1 = \frac{n}{n'}; V\left(\bar{x}'\right) = \lambda_3 S_{1x}^2 + \sum_{i=1}^{N} \lambda_{3i} S_{2xi}^2; Cov\left(P^*, \bar{x}'\right) = \lambda_3 S_{1xy} + \sum_{i=1}^{N} \lambda_{3i} S_{2xyi}^2 ;$$

where $\lambda_3 = \frac{1-f'}{n'}; f' = \frac{n'}{N}$; $\lambda_{3i} = \frac{M_i^2\left(1-f_{2i}\right)}{n'Nm_i}$; $P^* = \bar{x}^*$, $\bar{y}^*$, $\bar{x}'$; $Q^* = \bar{x}^*$, $\bar{y}^*$ and

$P^* \neq Q^*$.

The large sample approximation to the MSE of $\bar{y}_3^*$ is given by

$$MSE\left(\bar{y}_3^*\right) = V\left(\bar{y}^*\right) + B^2 V\left(\bar{x}' - \bar{x}^*\right) + 2BCov\left(\bar{y}^*, \bar{x}' - \bar{x}^*\right)$$

$$= V(\bar{y}^*) + B^2[V(\bar{x}^*) + V(\bar{x}^{'}) - 2B\,\mathrm{cov}(\bar{x}^{'}, \bar{x}^*)] + 2B[Cov(y^*) - Cov(\bar{y}^*) - Cov(\bar{y}^*, \bar{x}^*)]$$

If the values of the variance and covariance terms defined in section 2 are correspondingly substituted into (11), the MSE of $\bar{y}_3^*$ after simplification becomes

$$MSE\left(\bar{y}_3^*\right) = \lambda' S_{1B}^2 + \sum_{i=1}^{n'} \lambda_{2i} S_{2Bi}^2 - \lambda_3 S_{1B'}^2 - \sum_{i=1}^{n'} \lambda_{3i} S_{2Bi'}^2 \quad (12)$$

where $S_{1B}^2 = B^2 S_{1x}^2 - 2BS_{1xy}$, $S_{2Bi}^2 = B^2 S_{2xi}^2 - 2BS_{2xyi}$

Remark 1: it will be observed that the MSEs of the estimators $\bar{y}_1^*, \bar{y}_2^*$ and $\bar{y}_3^*$ each consist of four components. The first two components on the right hand sides of (8), (9) and (12) represents the additional contribution to the variance arising due to the fact that the values determined from the large sample of size n' are subject to error while the last two components represent the variance of the estimates if n' where equal to N. The estimators under consideration require the advance knowledge of some population parameters which are usually unknown. However in practice, the results from previous experience (survey) or the sample estimators of their population parameters may be substituted for this purpose. Although, the estimation may turn out to be biased the bias would be negligible in large samples and the approximate MSEs to order one will be the equivalent to those derived and for large samples, the difference would be minimal.

3. Efficiency Comparison

In this section, we considered the theoretical comparison of the performances of the suggested estimators with respect to Sukhatme et al (1984) estimators. It is easily seen that when we subtract (1), (2) and (3) from (8), (9) and (12) respectively we get:

Ratio estimation:

$$\lambda_1 \leq \lambda_3 S_{1R}^2 + \sum_{i=1}^{n} \lambda_{3i} S_{2Ri}^2$$

Product estimation:

$$\lambda_1 \leq \lambda_3 S_{1P}^2 + \sum_{i=1}^{n} \lambda_{3i} S_{2pi}^2$$

Regression estimation:

$$\lambda_1 \leq \lambda_3 S_{1B}^2 + \sum_{i=1}^{n} \lambda_{3i} S_{2Bi}^2$$

where $\lambda_1 = \left(\frac{\theta_1 - \theta}{n}\right) S_{1R}^2$; $\lambda_2 = \left(\frac{\theta_1 - \theta}{n}\right) S_{1B}^2$

Thus, we see that the estimators $\bar{y}_1^*, \bar{y}_2^*$ and $\bar{y}_3^*$ will be more efficient than $\bar{y}_1$, $\bar{y}_2$ and $\bar{y}_3$ respectively, since λ_1 and λ_2 will always be negative.

4. Numerical Illustration

The preceding theoretical results shall now be illustrated with reference to the data given by Okafor (2002) pp 223-224. Taking N=92, n'=26 and replacing the population values of (1), (2), (3), (8), (9) and (12) by their estimate obtained from the sample, the MSE for the different estimators are presented in table 1.

Table 1: MSE for the different estimators

Estimator	Ratio		Product		Regression	
	$\bar{y}_1$	$\bar{y}_1^*$	$\bar{y}_2$	$\bar{y}_2^*$	$\bar{y}_3$	$\bar{y}_3^*$
MSE	92.64	84.55	102.63	93.00	40.13	40.11

5. Conclusion

Table 1 clearly indicates that $MSE(\bar{y}_1^*)$ is smaller than $MSE(\bar{y}_1)$, $MSE(\bar{y}_2^*)$ is smaller than $MSE(\bar{y}_2)$. Hence, $\bar{y}_1^*$ and $\bar{y}_2^*$ is certainly to be preferred in practice. Also we see that the $MSE(\bar{y}_3^*)$ is smaller than

$MSE(\bar{y}_3)$ although the difference is not appreciable. This is an expected result since the conditions given in section 3 is satisfied. Note that $\hat{R} = 1.21$ is substantially different from $\hat{\beta} = 0.1$ the regression coefficient, this means that the regression line does not pass through the origin; and this goes further to explain why the regression estimators are better than the ratio and product estimators. It is therefore concluded that, for this data set the suggested estimators are more efficient than Sukhatme et al (1984) estimators.

References

Chaudhuri, A. and Stenger, H. (1992): *Survey sampling theory and methods*. Mracel Decker, Inc. New York.
Cochran W.G. (1940): The estimation of the yields of the cereal experiments by sampling for the ratio of gain to total produce. *Journal of Agricultural Science*, 30, 262-275.
Cochran, W.G. (1977): *Sampling techniques*. 3rd edition. John Wiley, New York.
Hossain, M.I., and Ahmed, M. S. (2001): A class of predictive estimators in two-stage sampling using auxiliary information. *Information and Management Sciences*, 12, 1, 49-55.
Kadilar, C. and Cingi, H. (2004): Estimator of a population mean using two auxiliary variables in simple random sampling, *International Mathematical Journal*, 5, 357–367.
Kadilar, C. and Cingi, H. (2005): A new estimator using two auxiliary variables, *Applied Mathematics and Computation*, 162, 901–908.
Murthy, M.N. (1964): Product method of estimation, *Sankhya*, A, 26, 69-74.
Okafor, C.F. (2002): *Sample survey theory with applications*. Afro-Orbis Publications Ltd Nsukka.
Pradhan, B.K. (2005): A chain regression estimator in TPS using multi-auxiliary information. *Bulletin Malaysian Mathematical Sciences*, 28, 1, 81-86.
Robson, D.S. (1957): Application of multivariate polykays to the theory of unbiased ratio-type estimation. *Journal of American Statistical Association,* 52, 511-522.
Sahoo, L.N. and Panda, P. (1997): A class of estimators in two-stage samplings with varying probabilities. *South African Statistical Journal*, 31, 151–160.
Sahoo, L.N. and Panda, P. (1999): A class of estimators using auxiliary information in two-stage sampling. *Australian and New Zealand Journal of Statistics*, 41, 4, 405–410.
Samiuddin, M. and Hanif, M. (2007): Estimation of population mean in single phase and two phase sampling with or with out additional information. *Pakistan Journal of Statistics*, 23, 2, 99-118.
Scott, A. and Smith, T.M.P. (1969): Estimation in multistage sampling. *Journal of American Statistical Association*, 64, 830-840.
Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984): *Sampling theory of surveys with applications*. Iowa State University Press, USA.
Upadhyaya, L. N., and Singh, H. P. (1999): Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, 41, 627-636.